Biodiversity Data Journal 14: e186683
doi: 10.3897/BDJ.14.e186683

Research Article

Transcriptome characterisation, SSR marker development and genetic diversity analysis of the endangered species *Camellia cucphuongensis* Ninh & Rosmann using Illumina sequencing

Bei Cui[‡], Lian Wang[‡], Quoc Khang Tran[§], Thi Tuyet Xuan Bui[|], Thi Tuyen Phung[¶], Syed Noor Muhammad Shah[#], Hieu Yen Ha[¤], Van Thang Bui[«], Duy Dinh Vu[»]

‡ Jiangsu Vocational Institute of Architectural Technology, School of Ecological Engineering, Xuzhou 221000, Jiangsu, China
§ The Olympia schools, To Huu Street, Hanoi, Vietnam
| Institute of Biology, Vietnam Academy of Science and Technology, Hanoi, Vietnam
¶ Faculty of Forest Resources and Environmental Management, Vietnam National University of Forestry, Xuan Mai, Hanoi, Vietnam
# Department of Horticulture, Gomal University, Dera Ismail Khan, Pakistan
¤ Faculty of Biotechnology, Vietnam National University of Agriculture, Ngo Xuan Quang, Gia Lam, Hanoi, Vietnam, Hanoi, Vietnam
« College of Forestry Biotechnology, Vietnam National University of Forestry, Xuan Mai, Hanoi, Vietnam, Hanoi, Vietnam
» Joint Vietnam - Russia Tropical Science and Technology Research Center, 63 Nguyen Van Huyen, Nghia Do, Hanoi, Vietnam

Corresponding author: Lian Wang (10743@jsjzi.edu.cn), Duy Dinh Vu (duydinhvu87@gmail.com)

Academic editor: Avelinah Julius

Received: 27 Jan 2026 | Accepted: 18 Mar 2026 | Published: 31 Mar 2026

Citation: Cui B, Wang L, Tran QK, Bui TTX, Phung TT, Shah SNM, Ha HY, Bui VT, Vu DD (2026) Transcriptome characterisation, SSR marker development and genetic diversity analysis of the endangered species *Camellia cucphuongensis* Ninh & Rosmann using Illumina sequencing. Biodiversity Data Journal 14: e186683. https://doi.org/10.3897/BDJ.14.e186683

Abstract

Overharvesting for ornamental and medicinal purposes, combined with ongoing habitat loss and fragmentation in Vietnam, has severely threatened wild populations of *Camellia cucphuongensis*. Effective conservation and management of this species, therefore, require robust genomic resources and informative molecular markers to quantify genetic diversity and population structure. In this study, we generated the first transcriptome dataset for *C. cucphuongensis* and developed expressed sequence tag–simple

© Cui B et al. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

sequence repeat (EST-SSR) markers using Illumina HiSeq™ 4000 sequencing. A total of 13,600,954 clean reads were obtained (Q20 = 97.55%, Q30 = 93.11%, GC = 44.08%). De novo assembly produced 118,552 unigenes with a mean length of 541.2 bp and an N50 of 683 bp. Functional annotation revealed that 52,107 and 25,640 unigenes had significant matches in the Nr and Swiss-Prot databases, respectively. Additionally, 28,007 unigenes were assigned to Gene Ontology terms, 27,968 to KOG categories and 11,959 to 117 KEGG pathways. Mining for simple sequence repeats identified 9,661 EST-SSR loci. From 60 screened primer pairs, 11 polymorphic EST-SSR markers were validated and applied to 60 individuals from three natural populations. Genetic diversity was moderate (N_E = 2.17; *PIC* = 0.548; H_O = 0.46; H_E = 0.50), with most variation occurring within individuals (79%) and 11% amongst populations (F_{ST} = 0.113; *Nm* = 1.96). Principal coordinate analysis (PCoA), discriminant analysis of principal components (DAPC), STRUCTURE and neighbour-joining (NJ) analyses all indicated detectable population structuring, with population CP showing clearer differentiation relative to LH and TL. Collectively, these transcriptomic resources and EST-SSR markers provide practical tools for genetic monitoring and can support conservation strategies that emphasise habitat protection and maintenance of connectivity to mitigate genetic erosion in this endangered golden camellia.

Keywords

Camellia cucphuongensis, Illumina HiSeq 4000, transcriptome, EST-SSR, genetic diversity, population structure

Introduction

Camellia (Theaceae) is a genus of evergreen shrubs and small trees predominantly found in East and Southeast Asia (Zhao et al. 2017). Estimates of species richness within the genus range from approximately 120 species (Ming and Bartholomew 2007) to nearly 280 species (Gao et al. 2005), reflecting both its taxonomic complexity and ongoing discovery of new taxa. Species of *Camellia* have high economic and cultural value due to their ornamental flowers, caffeine and purine alkaloid content and the extensive use of their seeds for edible oil production and their leaves in the pharmaceutical industry. Vietnam is a centre of diversity for the genus, harbouring more than 75 species, many of which are native or endemic (Nguyen et al. 2020). Despite their economic importance, *Camellia* species are increasingly threatened by overexploitation and habitat loss (Xin et al. 2017). Rapid economic development over recent decades has led to substantial deforestation, forest fragmentation and degradation of natural habitats, severely impacting *Camellia* populations (Mao et al. 2025). As a result, population sizes have declined sharply, raising concerns about genetic erosion (Bosse and van Loon 2022). At present, at least 51 *Camellia* species are listed as threatened at both national and global levels (Beech et al. 2017), highlighting the urgent need for effective conservation strategies informed by genetic data.

Camellia cucphuongensis Ninh & Rosmann, a rare golden camellia species, is geographically restricted and highly endangered. It is known to occur in Cuc Phuong National Park (Ninh Binh Province) and in Tan Lac and Lac Thuy districts (Phu Tho Province), Vietnam (Tran 1998), with additional records reported from Yunnan, China (Gao et al. 2005). In Vietnam, the species is mainly confined to secondary forests, where it flowers annually from October to December. Natural populations of *C. cucphuongensis* have experienced severe declines due to habitat conversion for agriculture and urbanisation, as well as intensive harvesting of flower buds for commercial trade. As a result, the species has been assessed as Critically Endangered (CR) under the IUCN Red List Criteria (Orel et al. 2018). Despite its high conservation value, genomic resources for *C. cucphuongensis* remain extremely limited and no comprehensive genetic or population-level studies have been conducted to date.

Understanding the genetic diversity and population structure of endangered species is essential for designing effective conservation and management programmes (Wariss et al. 2025). Microsatellite markers, also known as simple sequence repeats (SSRs), are short tandemly repeated DNA motifs that are ubiquitous in eukaryotic genomes (Vieira et al. 2016). Variation in the number of repeat units generates high levels of allelic polymorphism, making SSRs particularly informative genetic markers. SSRs are characterised by codominant inheritance, high abundance and reproducibility (Vu et al. 2020, Pham et al. 2024, Li et al. 2024, Bui and Vu 2024, Ha et al. 2025, Phan et al. 2026). There are various molecular marker systems, but their use depends on the study's objectives and the target species. For example, Amplified Fragment Length Polymorphisms (AFLPs) and Random Amplified Polymorphic DNA (RAPDs) can produce many markers without prior genomic information. However, the markers are dominant and have limited potential in heterozygosity and population structure analyses (Mueller and Wolfenbarger 1999). Single Nucleotide Polymorphisms (SNPs) are abundant and co-dominant, but their development and genotyping can be expensive. The demand for non-model organisms without a reference genome often requires approaches such as reduced-representation sequencing (Davey et al. 2011). In contrast to other molecular markers, SSRs markers are often species-specific, rendering them especially suitable for analysing intraspecific genetic diversity and population structure (Yang et al. 2025). Consequently, SSRs have been widely applied in studies of plant genetic relationships, population differentiation and conservation genetics (Liu et al. 2013, Fan et al. 2025). A significant knowledge gap exists concerning the genetic basis of environmental adaptation and stress tolerance in *C. cucphuongensis*. Addressing this gap through transcriptome sequencing is a powerful and economical strategy to elucidate the underlying molecular mechanisms and to generate genome-wide markers suitable for population genetic analyses. A transcriptome comprises the complete set of RNA molecules, including messenger RNA (mRNA), ribosomal RNA (rRNA), transfer RNA (tRNA) and other non-coding RNAs, expressed in a cell or tissue under specific environmental conditions or developmental stages (Martínez-Campos et al. 2025). Advances in next-generation sequencing (NGS) technologies, particularly high-throughput RNA sequencing (RNA-Seq), have enabled the rapid generation of large-scale expressed sequence data for non-model plant species (Li et al. 2021).

Transcriptome-based approaches enable the efficient development of numerous unigene-derived SSRs (EST-SSRs), which are valuable tools for genome-wide analyses due to their combination of functional relevance and high polymorphism (Wang et al. 2011, Tan et al. 2013, Tong and Gao 2020, Parmar et al. 2022, Xie et al. 2025). In this study, we used the Illumina HiSeq™ 4000 platform to generate a comprehensive transcriptome dataset for *C. cucphuongensis*. From the assembled unigenes, we developed a large set of EST-SSR markers, which was then used to assess genetic diversity and population structure amongst natural populations in Vietnam. These genomic resources provide a valuable foundation for functional gene annotation, population genetics and evolutionary studies within *Camellia* and offer critical molecular tools to support the conservation and sustainable management of this endangered species.

Material and methods

Plant material and RNA extraction

Fresh leaf, stem and root tissues of *C. cucphuongensis* were collected from four wild naturally occurring individuals in Cuc Phuong National Park, Ninh Binh Province, Vietnam (105°35′57″E; 20°21′06″N) in April 2025. Immediately after collection, samples were frozen in liquid nitrogen in the field and transported to the Laboratory of Molecular Biology, Joint Vietnam–Russia Tropical Science and Technology Research Center, where they were stored at −80°C until RNA extraction. Total RNA was isolated separately from each tissue sample using a Plant RNA Kit (Omega Bio-tek, Inc.) following the manufacturer's instructions, with an additional DNase I treatment to remove genomic DNA contamination. RNA integrity was assessed by electrophoresis on 1% agarose gels and further evaluated using an Agilent 2100 Bioanalyzer (Agilent Technologies, CA, USA). RNA concentration and purity were determined using a NanoDrop ND-2000 spectrophotometer (NanoDrop Technologies, DE, USA). Equal quantities of high-quality RNA from the different tissues were pooled to construct a representative transcriptome for Illumina sequencing, to maximise the diversity of transcripts captured in the reference assembly.

Library construction and Illumina sequencing

The pooled RNA sample was submitted to Breeding Biotechnologies Co., Ltd. for mRNA library construction and sequencing on the Illumina HiSeq™ 4000 platform (Illumina, Inc., CA, USA). Library preparation was performed using the TruSeq™ RNA Sample Preparation Kit (Illumina, CA, USA) according to the manufacturer's protocol. Briefly, poly(A)+ mRNA was enriched from total RNA using magnetic oligo(dT) beads (Oligotex mRNA Kit) and fragmented into short segments using an RNA fragmentation kit. First-strand cDNA synthesis was carried out using random hexamer primers and reverse transcriptase (Invitrogen, CA, USA), followed by second-strand cDNA synthesis using DNA polymerase I to generate double-stranded cDNA. Residual RNA was removed

using RNase H, after which end repair was performed and a single adenine base was added to the 3′ ends of the cDNA fragments. Sequencing adapters were then ligated to the cDNA fragments. Adapter-ligated fragments were enriched by PCR amplification and PCR products were size-selected by electrophoresis on 2% agarose gels (Certified Low Range Ultra Agarose). Fragments of appropriate size were purified and sequenced on the Illumina HiSeq™ 4000 platform to generate paired-end reads.

Sequence data processing and de novo transcriptome assembly

Raw sequencing reads were subjected to quality control to remove adapter sequences, reads containing more than 5% ambiguous nucleotides (N) and reads with more than 20% low-quality bases (Phred quality score < 10). High-quality clean reads were retained for subsequent analyses. The filtered reads were de novo assembled into contigs using Trinity version 2.8.5 software with default parameters (Grabherr et al. 2011). To reduce redundancy, the assembled contigs were further clustered into non-redundant unigenes using TGICL version 2.1 (Pertea et al. 2003).

Functional annotation of unigenes

To infer putative functions, all assembled unigene sequences were compared against several public databases, including the NCBI non-redundant (Nr) protein database (Deng et al. 2006), Swiss-Prot (Apweiler 2004), Gene Ontology (GO) (Ashburner et al. 2000), EuKaryotic Orthologous Groups of proteins (KOG) (Koonin et al. 2004) and Kyoto Encyclopaedia of Genes and Genomes (KEGG) (Kanehisa 2004), using BLASTX with an E-value threshold of $\leq 1 \times 10^{-5}$. Protein domain annotation was performed using HMMER software to search against the Pfam database (Finn et al. 2009). GO annotation and functional classification of the unigenes into biological process, molecular function and cellular component categories were conducted using Blast2GO version 5.2 software (Conesa et al. 2005). KOG classification was performed using BLAST searches against the KOG database and KEGG pathway mapping was used to identify metabolic pathways and associated gene functions represented in the transcriptome.

Development of EST-SSR markers

Microsatellite loci were identified from unigene sequences longer than 1 kb using MISA software (Beier et al. 2017). The minimum repeat thresholds were set as follows: mononucleotide repeats ≥ 12, dinucleotide repeats ≥ 6 and tri-, tetra-, penta- and hexanucleotide repeats ≥ 5. Based on the identified SSR loci, a total of 4,730 EST-SSR primer pairs were designed using Primer3 version 3.0 (Kõressaar et al. 2018) with the following parameters: primer length of 18-24 bp (optimal 20 bp), annealing temperature of 55-65°C (optimal 60°C), GC content of 40-65% (optimal 50%) and expected PCR product size of 100-300 bp. All primers were synthesised by Breeding Biotechnologies Co., Ltd.

DNA extraction and SSR polymorphism validation

From the designed primer set, 60 EST-SSR primer pairs were randomly selected for initial PCR screening. Amongst these, 11 primer pairs successfully amplified polymorphic loci across 60 individuals of *C. cucphuongensis* sampled from three wild natural populations: Cuc Phuong (Ninh Binh Province), Tan Lac and Lac Thuy (Phu Tho Province) (Fig. 1, Table 1). Within each population, individuals were sampled at least 300 m apart to reduce the probability of collecting those closely related and to ensure that each sample represented an independent genetic individual. Genomic DNA was extracted from fresh leaf tissue using a Plant DNA Kit (BioTeke, Beijing, China) following the manufacturer's instructions. DNA quality and concentration were assessed using a NanoDrop ND-2000 spectrophotometer and normalised to 20 ng·μl^{-1}. PCR amplifications were performed in 25 μl reaction volumes containing 2.5 μl template DNA, 12.5 μl 2× Taq Master Mix, 1 μl of each primer and 8 μl nuclease-free water. Amplifications were conducted on a GeneAmp PCR System 9700 (Applied Biosystems, USA) under the following conditions: initial denaturation at 94°C for 3 min; 40 cycles of 94°C for 30 s, primer-specific annealing for 30 s and extension at 70°C for 1 min; followed by a final extension at 72°C for 10 min. PCR products were resolved on 8% polyacrylamide gels in TAE buffer and visualised using GelRed™ nucleic acid gel stain. Fragment sizes were estimated using GenoSens 850 gel analysis software (Clinx Science Instruments Co., Ltd.) with a 35 bp DNA ladder (Invitrogen). Amplicon sizing and quantification were further verified using an Agilent 5300 Fragment Analyzer with a DNF-905 dsDNA kit (Agilent, USA).

Figure 1.

Morphological characteristics of *C. cucphuongensis* Ninh & Rosmann: **A** adult plant in natural habitat; **B** flowering branch; **C** flower in full bloom; **D** flower bud; **E** fruits and leaves; **F** seeds (Photo: Dr. Thi Tuyen Phung).

Table 1.

Sampling locations and genetic diversity parameters of *C. cucphuongensis* populations, based on 11 EST-SSR loci.

Population code	Locations	Latitude (N), Longitude (E)	N	N_A	N_E	P(%)	H_O	H_E	F_{IS}	$F_{IS}IIM$
CP	Cuc Phuong National Park, Ninh Binh Province	20°21′06″, 105°35′57″	21	3.73	2.40	100	0.49	0.53	0.04	0.012
LH	Lac Thuy commune, Phu Tho Province	20°30'25", 105°43'51"	17	4.18	2.24	100	0.46	0.54	0.16	0.020
TL	Tan Lac commune, Phu Tho Province	20°33′29″, 105°20′10″	22	3.82	1.87	100	0.43	0.44	0.01	0.013
Species level			60	3.91	2.17	100	0.46	0.50	0.07	0.015

Note: N_A, mean number of alleles per locus; N_E, mean number of effective alleles; P(%), percentage of polymorphic loci; H_O and H_E, mean observed and expected heterozygosities, respectively; F_{IS}, inbreeding coefficient with *p < 0.05. **P < 0.01, ***P < 0.001; ns, not significant; The individual inbreeding model was performed to evaluate the F_{IS} index for null allele frequency $F_{IS}IIM$.

Genetic diversity analyses

Potential genotyping errors and null alleles were assessed using Micro-Checker version 2.0 (van Oosterhout et al. 2004). Polymorphism information content (*PIC*) values were calculated using CERVUS software (Kalinowski et al. 2007). Standard genetic diversity parameters, including number of alleles per locus (N_A), effective number of alleles (N_E), observed heterozygosity (H_O), expected heterozygosity (H_E), inbreeding coefficient (F_{IS}) and gene flow (*Nm*), were estimated using GenAlEx version 6.5 (Peakall and Smouse 2012). Deviations from Hardy-Weinberg equilibrium (*HWE*) were tested in CERVUS with 1,000 permutations.

Population structure and phylogenetic analyses

Genetic differentiation amongst populations was quantified using Wright's fixation index (F_{ST}) (Weir and Cockerham 1984) and G'ST (Hedrick 2005), as implemented in GenAlEx 6.5. The significance of pairwise F_{ST} values was evaluated using Alequin version 3.1 with 10,000 permutations (Excoffier et al. 2005). Analysis of molecular variance (AMOVA) was also performed in Alequin to partition genetic variation within and amongst populations. Bayesian clustering analysis was conducted using Structure version 2.3.4 (Pritchard et al. 2000) under an admixture model with correlated allele frequencies. Ten independent runs were performed for each value of K (1-15), with 500,000 Markov Chain Monte Carlo

(MCMC) iterations following a burn-in period of 100,000 iterations. The optimal number of clusters was determined using Structure Harvester version 0.6.93 (Earl and vonHoldt 2011), based on the ΔK method (Evanno et al. 2005). Results from replicate runs were aligned using Clumpp version 1.1.2 (Jakobsson and Rosenberg 2007) and graphical outputs were generated using Distruct version 1.1 (Rosenberg 2003). Discriminant analysis of principal components (DAPC) was performed using the adegenet package in R version 4.0.2 (Jombart et al. 2010) to identify genetic clusters without prior population information. The optimal number of clusters was determined using the Bayesian Information Criterion (BIC), with K values ranging from 1 to 20. Cross-validation (xvalDAPC) was used to retain the first 14 principal components, explaining 98.5% of the total variance and seven discriminant functions. Phylogenetic relationships amongst individuals were inferred using the neighbour-joining (NJ) method, based on Nei's genetic distances, implemented in Ntsys version 2.11 (Saitou and Nei 1987) and Mega version 11.0 (Tamura et al. 2021).

Results

Illumina sequencing and de novo assembly

Illumina transcriptome sequencing of *C. cucphuongensis* generated a total of 17,379,703 clean reads. After quality filtering, 13,600,954 clean reads were retained, with Q20 = 97.55% and Q30 = 93.11%, respectively. De novo assembly was performed using Trinity, resulting in 1,731,255 contigs with a total length of 166,947,590 bp, an N50 of 155 bp and
a mean contig length of 964.3 bp. The majority of contigs (95.41%) were between 0–300 bp in length, while 0.59% ranged from 1,001–2,000 bp. A total of 4,468 contigs exceeded 2,001 bp. From these, 197,570 transcripts were obtained, with a total length of 179,788,275 bp, an N50 of 1,675 bp and a mean length of 910 bp. Amongst these, 24,334 transcripts were longer than 2,001 bp. After redundancy reduction, 118,552 unigenes were generated, with a total length of 64,158,962 bp, an N50 of 683 bp and a mean length of 541.2 bp. Most unigenes fell within the 200-300 bp (44.33%) and 301–500 bp (28.97%) ranges, while 3.82% exceeded 2,001 bp (Table 2).

Table 2.
Summary statistics of de novo transcriptome assembly for *C. cucphuongensis*.

Length range (bp)	Unigene	Contigs	Transcripts
0-300	52,559	1,651,729	59,627
301-500	34,347	41,773	44,156
501-1000	18,366	23,076	35,986
1001-2000	8,747	10,209	33,467
> 2000	4,533	4,468	24,334
Total Number	118,552	1,731,255	197,570

Length range (bp)	Unigene	Contigs	Transcripts
Total Length	64,158,962	16,6947,590	179,788,275
N50 Length	683	155	1,675
Mean Length	541.2	1,651,729	59,627

Functional annotation and classification of unigenes

A total of 52,107, 31,987, 25,640, 28,007, 27,968 and 11,959 unigenes were annotated against the Nr, Pfam, Swiss-Prot, GO, KOG and KEGG databases, respectively (Table 3). Species distribution, based on the Nr database (Suppl. material 1), showed that the top matches were *Vitis vinifera* (7,134), *Dothistroma septosporum* (1,823), *Coffea canephora* (1,445), *Theobroma cacao* (1,317), *Nelumbo nucifera* (1,211), *Sesamum indicum* (1,149), *Ziziphus jujuba* (1,012), *Manihot esculenta* (890), *Citrus sinensis* (748) and *Jatropha curcas* (744). An additional 34,576 unigenes showed similarity to a broad range of other species. GO annotation assigned 28,007 unigenes (23.62%) to three ontologies comprising 51 functional terms (Fig. 2). The biological process category represented the largest proportion (51,939; 42.37%), followed by cellular component (36,389; 29.69%) and molecular function (34,254; 27.94%). Within the cellular component ontology, the most frequently represented terms were cell, cell part, organelle and membrane. In the molecular function ontology, catalytic activity and binding predominated, while, in the biological process, metabolic process and cellular process were the dominant terms. KOG classification assigned 27,968 unigenes to 25 functional categories (Fig. 3). The most abundant category was general function prediction only (5,697; 20.37%), followed by post-translational modification, protein turnover and chaperones (2,912; 10.41%), translation, ribosomal structure and biogenesis (2,263; 8.09%) and signal transduction mechanisms (1,865; 6.67%). KEGG mapping assigned 11,959 unigenes (10.09%) to 117 pathways across five major functional classes (Fig. 4). Metabolism was the largest category (3,860; 50.92%), followed by genetic information processing (2,494; 32.9%), cellular processes (498; 6.57%), organismal systems (383; 5.05%) and environmental information processing (345; 4.55%).

Frequency and distribution of EST-SSRs

A total of 13,280 sequences (total length 25,815,979 bp) were examined for SSR discovery. In total, 9,661 SSRs were identified; 733 were compound SSRs and 2,261 unigenes contained more than one SSR locus (Table 4). A total of 6,558 sequences contained SSRs. The overall microsatellite frequency in unigenes was 72.75%, with an average density of one SSR per 2.67 kb. SSRs were classified into six repeat types (Suppl. material 2): mononucleotide repeats (3,986; 41.26%) and dinucleotide repeats (3,926; 40.64%) were the most abundant, followed by trinucleotides (1,617; 16.74%), tetranucleotides (102; 1.06%), pentanucleotides (13; 0.13%) and hexanucleotides (17; 0.18%). Across repeat numbers, SSRs with 10 repeats (1,662; 17.2%) were most frequent, followed by those with six repeats (1,415; 14.65%) and five repeats (1,000; 10.35%), whereas each of the remaining repeat number classes each represented <

10% of SSRs (Suppl. material 4). Motif composition showed that, amongst mononucleotides, A/T (3,908; 40.45% of total SSRs) dominated, with C/G (78; 0.81%) being rare. Amongst dinucleotides, AG/CT (3,025; 31.31%) was most abundant, followed by AT/AT (655; 6.78%) and AC/GT (244; 2.53%). Amongst trinucleotides, AAG/CTT (415; 4.3%) was the most frequent motif. Amongst tetranucleotides, AAAT/ATTT (30; 0.31%) was the predominant motif, followed by AAAG/CTTT (19; 0.2%). Overall, the 17 most frequent motifs accounted for 99.05% of all SSRs (Suppl. material 3).

Table 3.

Functional annotation of *C. cucphuongensis* in different databases.

Annotated database	Annotated	300–1000 (bp)	≥ 1000(bp)
COG	17,454	6,607	4,696
GO	28,007	11,717	6,703
KEGG	11,959	5,192	2,881
KOG	27,968	11,923	7,350
Pfam	31,987	13,156	9,943
Swissprot	25,640	11,332	7,797
Nr	52,107	22,648	11,811
All	53,845	23,182	11,836

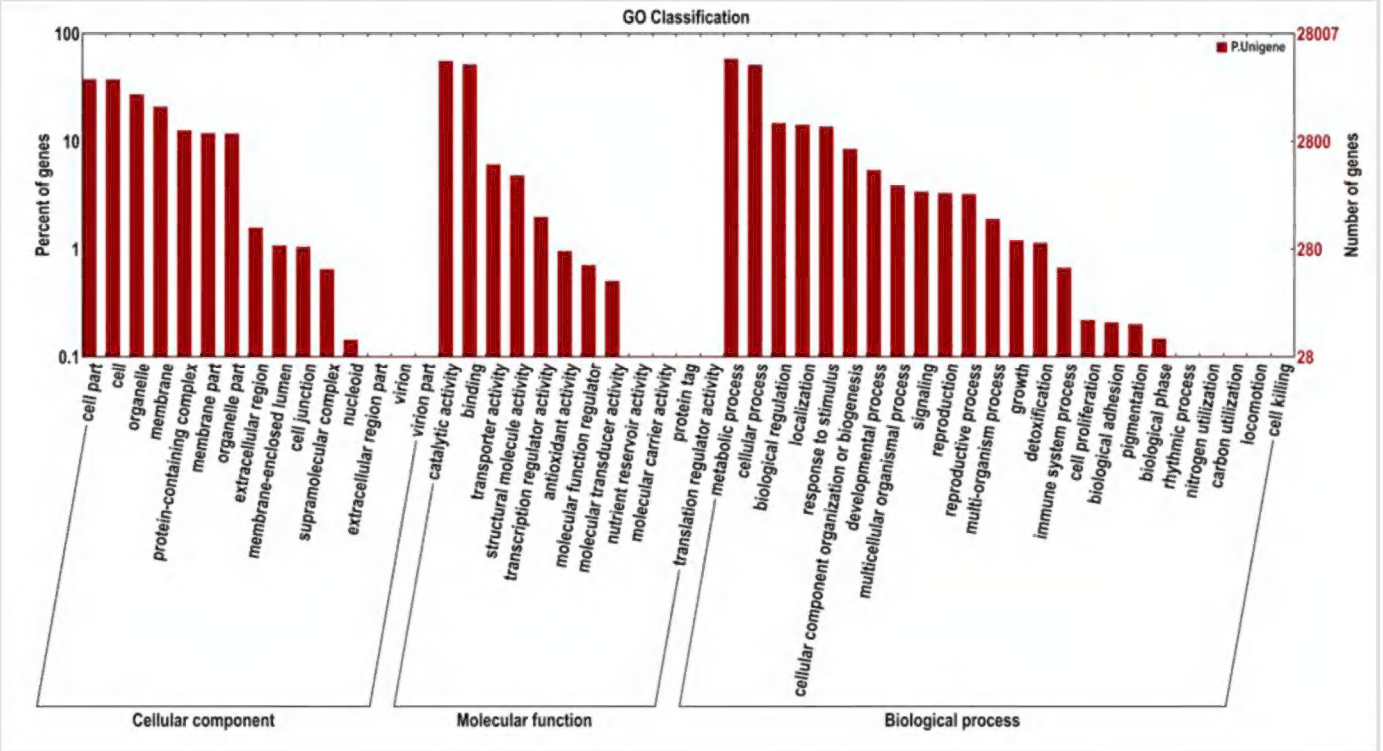

Figure 2.

Gene Ontology (GO) classification of *C. cucphuongensis* unigenes into three main categories: biological process, molecular function and cellular component.

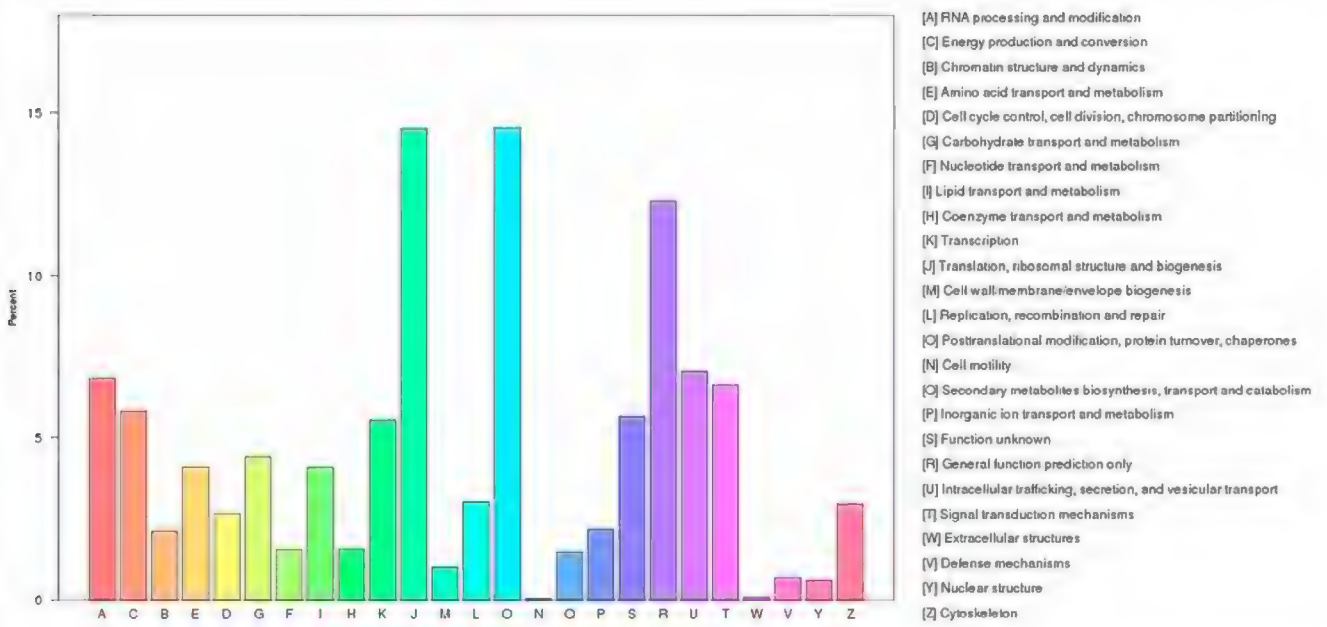

Figure 3.

Functional classification of *C. cucphuongensis* unigenes, based on Eukaryotic Orthologous Groups (KOG).

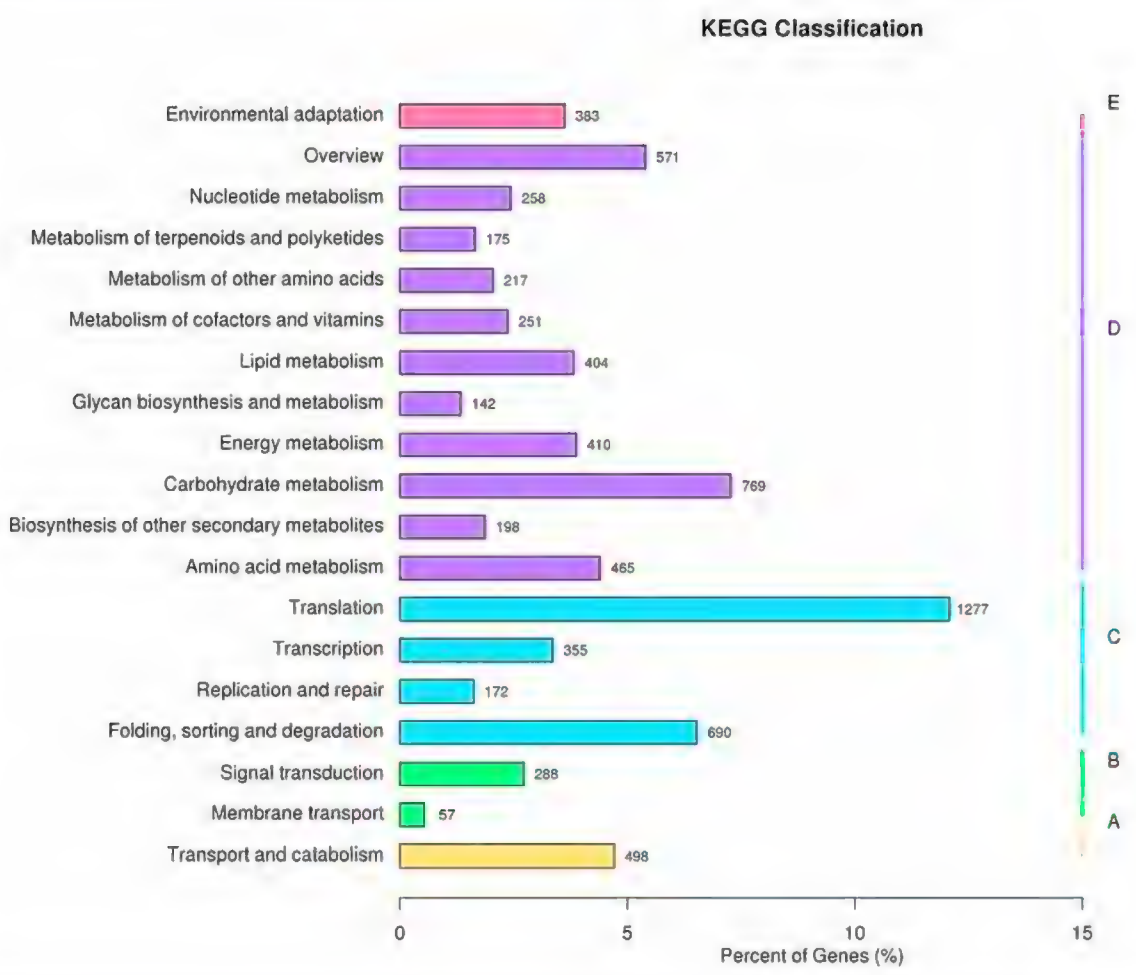

Figure 4.

Functional classification of *C. cucphuongensis* unigenes,based on Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathway annotations: Cellular Processes (A); Environmental Information Processing (B); Genetic Information Processing (C); Metabolism (D); Organismal Systems (E).

Table 4.

Summary of analyses of expressed sequence Tag–Simple Sequence repeat (EST-SSRs) in *C. cucphuongensis*.

Item	Parameters	Number
EST-SSR	Total number of sequences examined	13,280
	Total size of examined sequences (bp)	25,815,979
	Total number of identified SSRs	9,661
	Number of SSR containing sequences	6,558
	Number of sequences containing more than 1 SSR	2,261
	Number of SSRs present in compound formation	733

Validation of EST-SSR markers and genetic diversity

Of the 4,730 designed EST-SSR primer pairs, 60 were initially screened across 10 individuals. This identified 11 polymorphic loci and 32 monomorphic loci. The 11 robust polymorphic markers were then used for population genetic analyses. At the species level (N = 60), these 11 loci yielded mean values of N_A = 3.91, N_E = 2.17, *PIC* = 0.548, H_O = 0.46 and H_E = 0.50, with a mean F_{IS} = 0.07 and F_{IS}*IIM* = 0.015 (Tables 1, 5). All three populations showed 100% polymorphic loci (Table 1). Population-level genetic diversity parameters were as follows: CP (N_A = 3.73, N_E = 2.40, H_O = 0.49 and H_E = 0.53), LH (N_A = 4.18, N_E = 2.24, H_O = 0.46 and H_E = 0.54) and TL (N_A = 3.82, N_E = 1.87, H_O = 0.43 and H_E = 0.44). Population-level inbreeding coefficients (F_{IS}) were 0.04 (CP), 0.16 (LH), and 0.01 (TL), indicating heterozygote deficiency particularly in LH (Table 1).

Table 5.

Genetic parameters in eleven SSR loci for *C. cucphuongensis*.

Locus	Sequence of primer (5'–3')	Motif type	Size (bp)	N_A	N_E	Null allele	*PIC*	H_O	H_E	F_{IS}	F_{IT}	F_{ST}	P_{HWE}
SSR01	F: AGA TAA AAG ACC CAC GTC AAG G R: GCA ACT CCA AAA CAC ACT CAT T	(TA)9	126-144	5.00	2.04	No	0.909	0.48	0.52	-0.09	-0.04	0.05	ND
SSR 02	F: AAT ACG AAC ACG TTG AGA GCA A R: TTC AGT TCC AAT CCA AGC G	(CAC)6	252-264	3.67	2.09	Yes	0.488	0.50	0.20	0.59	0.61	0.05	NS
SSR03	F: ACA GCC AAA ATC AAT CCC TAG A R: ACA ACA CAA CAC AAC ACA GTC G	(AG)7	262-298	4.33	2.79	Yes	0.493	0.64	0.32	0.50	0.55	0.09	***
SSR04	F: ACT ACC TCC AGC AAT GGA TCT C R: GAC CGT GTA ATT CGG GTT CTT A	(CT)6	192-220	4.67	2.50	No	0.680	0.58	0.62	-0.06	-0.01	0.05	*

Locus	Sequence of primer (5'–3')	Motif type	Size (bp)	N_A	N_E	Null allele	*PIC*	H_O	H_E	F_{IS}	F_{IT}	F_{ST}	P_{HWE}
SSR05	F: AAA AGA GAG AGA GGA TGG ACC C R: TAA CCC TTC AAT TTC CCT TCA C	(AG)6	198-216	3.00	1.63	No	0.570	0.33	0.40	-0.23	-0.13	0.08	NS
SSR06	F: TTA TCT TCC TCA AGC TCT CCC TC R: GTT GAA ACG CAG AGG TGG TAA	(CT)8	200-216	2.67	1.80	No	0.334	0.41	0.43	-0.05	0.10	0.14	ND
SSR07	F: TCT CTC TCT CTC GGT TTA GGG TT R: GCT TGC TTG ATC TCC TTG ATG	(CT)7	238-259	4.67	2.02	Yes	0.437	0.49	0.42	0.13	0.28	0.18	NS
SSR08	F: CCT CTC CTC CCT CCT CCT T R: GAC AAA CAA AGA ATC CAA AGC C	(TCC)7	267-282	3.67	2.28	Yes	0.549	0.50	0.42	0.16	0.26	0.11	*
SSR09	F: CTG GAA AAC TGT CAA AAC CAA C R: TGC CTC TTT TAT ATC TGT CTC CTT C	(AG)7	182-202	4.00	2.39	No	0.538	0.56	0.62	-0.10	0.05	0.13	NS
SSR10	F: CAG CCA TTG ATG TTG ATG ACT T R: TCC AAA ACC AGA AAG CCT AGA A	(TC)8	146-170	3.67	1.87	No	0.590	0.45	0.47	-0.04	0.04	0.08	NS
SSR11	F: TCAACATGGAACTCCCACAA R: TGATGCTTTGTGATGGATGG	(CAG)10	196-258	3.67	2.46	No	0.435	0.59	0.63	-0.07	0.02	0.08	NS
Mean				**3.91**	**2.17**		**0.548**	**0.46**	**0.5**	**0.07**	**0.16**	**0.09**	

Note: number of alleles (N_A), effective alleles (N_E), polymorphism information content (*PIC*), observed heterozygosity (H_O), expected heterozygosity (*HE*), inbreeding coefficient (F_{IS}), Null allele the average null allele frequency, coefficient of total inbreeding (F_{IT}), genetic differentiation index of Weir and Cockerham (F_{ST}), ns=not significant, *P < 0.05, **P < 0.01, ***P < 0.001

Population differentiation, clustering and phylogenetic patterns

The analysis of molecular variance (AMOVA: 9,999 permutations) revealed that the majority of genetic variation was distributed within individuals (79%), with 11% amongst populations and 10% amongst individuals within populations (Table 6). The F_{ST} value was 0.113 ($P < 0.05$), corresponding to a gene flow (*Nm*) estimate of 1.96 (Table 6). Pairwise comparisons indicated moderate genetic structure amongst the populations. The F_{ST} values were 0.114 between CP and LH, 0.138 between CP and TL and 0.081 between LH and TL, with corresponding *N*m values of 1.941, 1.568 and 2.843, respectively (Suppl. material 5). Principal Coordinate Analysis (PCoA) separated individuals primarily along the first axes, which cumulatively explained 64.01% of the variance (Fig. 5A). Discriminant Analysis of Principal Components (DAPC) without prior population information identified three genetic clusters corresponding to geographic origins CP, LH and TL; DAPC incorporating prior information further illustrated assignment patterns within and amongst populations (Fig. 5B and C). Bayesian clustering analysis implemented in STRUCTURE supported $K = 2$ as the most likely

number of genetic clusters (ΔK = 312.8). All individuals exhibited admixture between two ancestral components (Fig. 5E). Cluster 1 was most frequent in CP (82.0%), moderate in LH (53.2%) and lowest in TL (20.4%), whereas Cluster 2 showed the opposite pattern, contributing 18.0%, 46.5% and 79.6% in CP, LH and TL, respectively (Fig. 5E). A neighbour-joining tree, based on Nei's genetic distance, revealed two major branches, but the relationships amongst individuals were not strictly concordant with geographic origin (Fig. 5D). This pattern is consistent with the admixture proportions inferred from STRUCTURE and the overlap observed in ordination analyses.

Table 6.

Analysis of molecular variance from natural populations for *C. cucphuongensis*.

	df	Sum of squares	Variance components	Total variation (%)	F-statistics
Amongst populations	2	35.081	0.362	11	F_{ST} = 0.113*** F_{IS} = 0.112*** F_{IT} = 0.213*** Nm = 1.96
Amongst individuals within populations	57	179.644	0.317	10	
Within individuals	60	151.000	2.517	79	
Total	119	365.725	3.196	100	

Note: coefficient within population (F_{IS}); coefficient of the total population (F_{IT}); coefficient of differentiation (F_{ST}); Gene flow (Nm). df: Degrees of freedom; *P < 0.05, **P < 0.01, ***p < 0.001.

Discussion

Transcriptome assembly quality and comparative evaluation within *Camellia*

Microsatellites or simple sequence repeats (SSRs) are tandemly repeated motifs of 1–6 nucleotides that are ubiquitously distributed throughout eukaryotic genomes, occurring in both coding and non-coding regions (Tautz and Renz 1984, Charlesworth et al. 1994). Owing to their high levels of polymorphism, codominant inheritance and reproducibility, SSR markers have been widely applied in studies of genetic structure, gene mapping, evolutionary biology and plant breeding (Kumar et al. 2024). In addition to their utility as molecular markers, SSRs may also influence gene regulation and function, thereby contributing to phenotypic variation and adaptive evolution (Ahmad et al. 2018). In the present study, we report the first comprehensive transcriptomic characterisation of *C. cucphuongensis* using next-generation sequencing (NGS). The Illumina HiSeq™ 4000 platform generated 118,552 unigenes, with an N50 length of 683 bp and a mean unigene length of 541 bp, providing a substantial genomic resource for this endangered species. Although these assembly metrics are lower than those reported for *Camellia taliensis* (N50 = 995 bp; mean = 685 bp) (Zhang et al. 2015), *C. oleifera* (N50 = 962 bp; mean = 708 bp) and *C. oleifera* (N50 = 771 bp) (Xia et al. 2014), they exceed those reported for *C. sinensis* (N50 = 480 bp; mean = 402 bp) (Tan et al. 2013). Our N50 is moderate

compared to some *Camellia* species, but it is higher than that of *C. sinensis* (480 bp). This shows that high-quality transcriptomic resources can be produced even for rare species with limited samples. The N50 value has a direct influence on the capture ability of full-length transcripts and identifies functional domains. For conservation genomics applications, this assembly quality provides sufficient resolution for marker development and diversity assessment, though it may limit full-length gene discovery. Future conservation efforts that need full gene characterisation would benefit from long-read or deeper sequencing technologies.

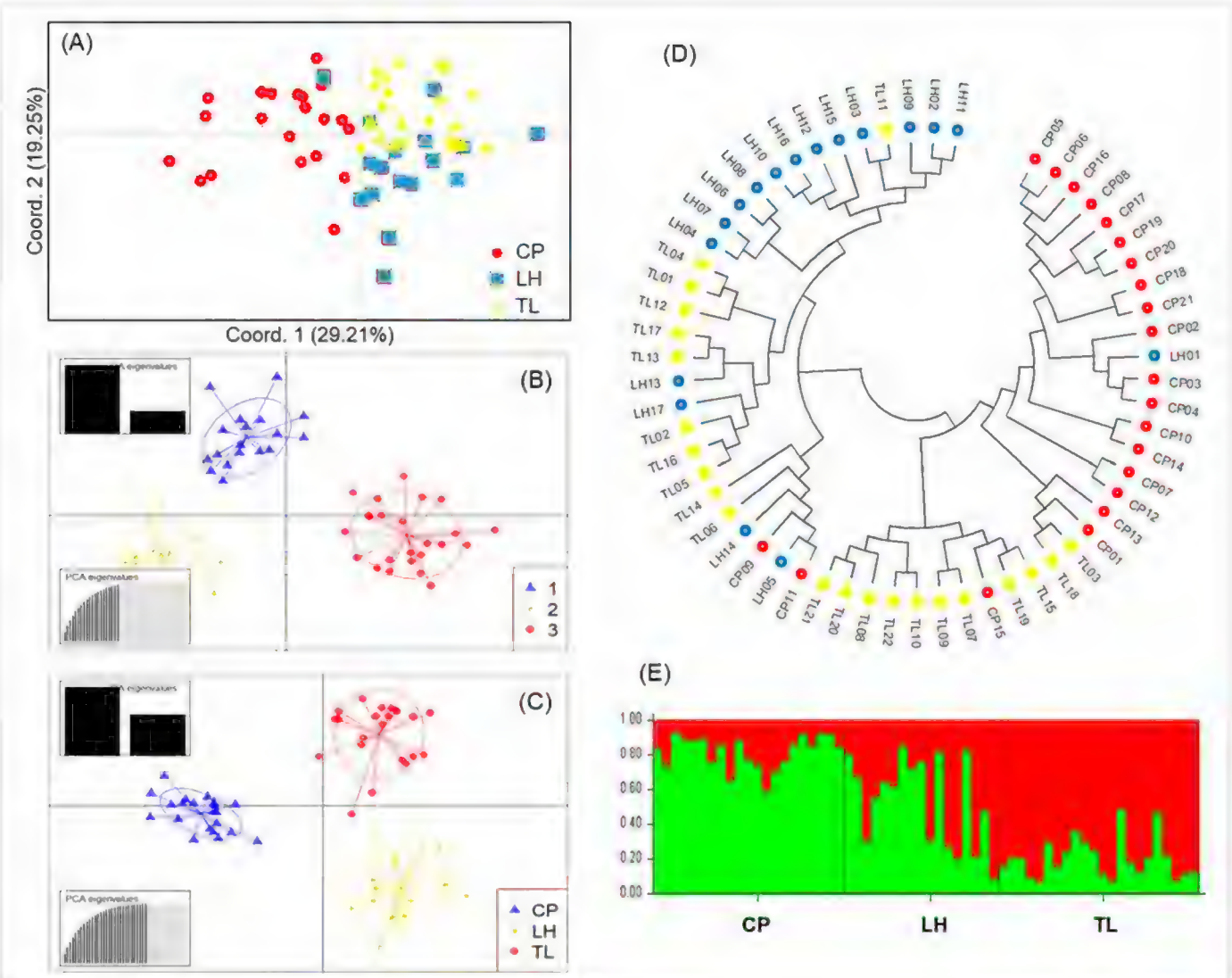

Figure 5.

Genetic relationships and population structure of *C. cucphuongensis* inferred from 11 EST-SSR loci. **(A)** Principal coordinate analysis (PCoA) of individuals from three populations (CP, LH, TL); **(B)** DAPC scatter plot inferred without prior population information; **(C)** DAPC scatter plot using prior population information; **(D)** Neighbour-Joining tree based on Nei's genetic distance; **(E)** Bayesian clustering analysis (STRUCTURE) showing the inferred genetic clusters across populations.

Functional annotation reveals conserved metabolic and regulatory profiles in *Camellia*

Functional annotation against multiple public databases (Nr, Pfam, Swiss-Prot, GO, KOG and KEGG) provided extensive insights into the functional landscape of the *C.*

cucphuongensis transcriptome. GO classification revealed that metabolic and cellular processes were the most abundant terms in the biological process category, while cell and cell part dominated the cellular component category. In the molecular function category, catalytic activity and binding were predominant. This functional distribution aligns with transcriptomic profiles reported for other *Camellia* species, including *C. oleifera* (Ye et al. 2020), *C. chekiangoleosa* (Tian et al. 2022), *C. japonica* (Li et al. 2021) and *C. sinensis* (Parmar et al. 2022). KOG annotation further revealed that the largest functional classes were "general function prediction only" and "post-translational modification, protein turnover and chaperones", closely reflecting patterns observed in *C. taliensis* (Zhang et al. 2015) and *C. oleifera* (Ye et al. 2020). These findings suggest that the transcriptomic architecture of *C. cucphuongensis* is broadly conserved within the genus, with a substantial proportion of genes involved in core cellular maintenance and regulatory processes. The over-representation of "general function prediction only" in KOG classification (20.37%) highlights substantial portions of the transcriptome with unknown functions, which is a common pattern in non-model species. For conservation, this highlights the importance of protecting the species' entire habitat to preserve unknown, but potentially adaptive genetic elements. KEGG pathway analysis highlighted the predominance of pathways related to ribosome functions, oxidative phosphorylation, RNA transport, purine metabolism, glycolysis/gluconeogenesis and the spliceosome, indicating robust transcriptional and metabolic activity. As the metabolic pathway representation is 50.92% of KEGG annotations, it suggests that *C. cucphuongensis* retains full metabolic capacity, despite population decline. This is an encouraging sign for conservation as it shows that genetic erosion has not yet compromised essential physiological functions. Collectively, these results demonstrate that the *C. cucphuongensis* transcriptome encompasses a wide array of molecular functions and biological pathways, providing a strong foundation for understanding the genetic basis of development, metabolism and stress responses in this endangered species. By identifying individuals or populations with unique alleles in these critical pathways, conservation managers can prioritise them for *ex situ* collection or habitat protection, ensuring that the maximum adaptive potential of the species is preserved.

EST-SSR characteristics and comparison with other *Camellia* species

A total of 9,661 EST-SSR loci were identified from the *C. cucphuongensis* transcriptome, providing the first SSR resource developed for this species. Mononucleotide and dinucleotide repeats were the most abundant SSR types, followed by trinucleotide repeats. This distribution pattern consistent with those reported for *C. chekiangoleosa* (Tian et al. 2022), *C. japonica* (Li et al. 2021) and *C. sinensis* (Parmar et al. 2022). The strong predominance of A/T-rich motifs, particularly the high frequency of A/T mononucleotide repeats and AG/CT dinucleotide repeats, reflects a significant AT bias in expressed regions, a characteristic widely observed in plant transcriptomes. In contrast, CG/CG motifs were rare in *C. cucphuongensis*, being consistent with previous findings in other *Camellia* species and other angiosperms, likely due to methylation-associated mutation biases in CG dinucleotides. The GC content of *C. cucphuongensis* unigenes (44.08%) was lower than that reported for *C. sinensis* (48%) (Tan et al. 2013), but

comparable to values observed in *C. oleifera* (41.51%) (Ye et al. 2020) and *C. oleifera* (39.1%) (Xia et al. 2014). GC content is an important genomic feature influencing DNA stability, gene expression and codon usage bias and interspecific differences may be associated with long-term evolutionary adaptation to distinct ecological niches. The dominance of AAG/CTT amongst trinucleotide repeats and AAAT/ATTT amongst tetranucleotide repeats further supports the notion that SSR motif composition is highly conserved across the genus, suggesting shared evolutionary constraints on microsatellite formation and maintenance.

Genetic diversity, population structure and evolutionary implications

Using 11 polymorphic EST-SSR markers, moderate genetic diversity was detected at the species level *C. cucphuongensis* (N_E = 2.17, H_O = 0.46, H_E = 0.50). These values are lower than those reported for more widely distributed or less disturbed *Camellia* species, such as *C. taliensis* (N_E = 6.6, H_O = 0.79, H_E = 0.82) (Wang et al. 2023), but comparable to those observed in *C. chekiangoleosa* (Tian et al. 2022), *C. nitidissima* (Li et al. 2019) and *C. nitidissima* var. *phaeopubisperma* (Xie et al. 2025). The reduced genetic diversity in *C. cucphuongensis* likely reflects its restricted geographic distribution, small population sizes and long-term anthropogenic pressures, such as habitat loss and overharvesting. AMOVA revealed that 79% of genetic variation was partitioned within individuals, with only 11% occurring amongst populations. This is pattern is typical of perennial, outcrossing woody species (Hamrick and Godt 1990) and is consistent with findings in *C. taliensis* (Wang et al. 2023) and *C. nitidissima* (Li et al. 2019). The moderate overall F_{ST} (0.09-0.11) and relatively high estimated gene flow (*Nm* = 1.96) suggest that genetic drift has not yet resulted in strong population differentiation, despite spatial separation amongst populations. This gene flow may be facilitated by life-history traits of *Camellia*, including outcrossing mating systems, long lifespan, overlapping generations and effective pollen dispersal. Nevertheless, positive F_{IS} values across populations indicate heterozygote deficiency, suggesting inbreeding or mating amongst related individuals. This is likely driven by population fragmentation and reduced effective population sizes.

Population structure analyses and conservation implications

Multivariate (PCoA, DAPC) and Bayesian (STRUCTURE) analyses consistently revealed genetic structuring amongst the three populations (CP, LH and TL). CP formed a genetically distinct unit, while LH and TL showed closer genetic affinity. The partial overlap observed in PCoA and the clear separation in DAPC reflect differences in analytical sensitivity: DAPC being more effective at maximising amongst-group variation. From a conservation perspective, these findings indicate that CP should be considered a distinct management unit. The closer genetic affinity between LH and TL populations may be the influence of natural and anthropogenic factors. There is no direct evidence of transplantation. However, the long history of local activities in the region, for example, traditional uses of *Camellia* species, may raise the possibility that seed dispersal or modification in the habitat could have influenced current genetic patterns. However, sampling from wild individuals and a moderate gene flow (*Nm* = 2.84 between LH and

TL) suggests the gene flow through natural cross-pollination across these populations. It should still be managed separately to preserve local genetic variation. The moderate genetic diversity and ongoing gene flow suggest that *C. cucphuongensis* retains evolutionary potential; however, continued habitat degradation and overexploitation could rapidly erode this genetic reservoir.

Conclusions

This study provides the first transcriptome resource and EST-SSR marker set for the endangered species *C. cucphuongensis* generated using the Illumina HiSeq™ 4000 platform. A total of 118,552 unigenes were assembled (N50 683 bp, mean length 541 bp) and 9,661 EST-SSRs were identified, enabling the development of informative microsatellite markers for this species. Validation of transcriptome-derived markers yielded 11 polymorphic EST-SSR loci, which were successfully applied to assess genetic diversity and population structure across three wild populations (CP, LH and TL). Overall genetic diversity was moderate (N_E = 2.17; H_O = 0.46; H_E = 0.50; PIC = 0.548), with moderate differentiation amongst populations (F_{ST} = 0.113; Nm = 1.96) and most variation distributed within individuals (79%). Multivariate and Bayesian analyses supported population structuring, with CP showing greater differentiation relative to LH and TL. Collectively, the transcriptome dataset and EST-SSR markers reported here provide practical molecular tools for population monitoring and conservation genetics of *C. cucphuongensis*. Given ongoing habitat loss and overharvesting pressures, conservation actions should prioritise habitat protection and the maintenance of connectivity amongst populations to reduce further genetic erosion.

Acknowledgements

We thank Bioeditas Technology Corporation (Shanxi, China) for assistance with data analysis. We thank local people for their expertise and support for this research.

Funding program

This research was funded by Jiangsu Construction Vocational and Technical College High-level Talent Project (China) (90250824523033)

Author contributions

Conceptualisation, Bei Cui and Dinh Duy Vu; methodology, Bei Cui and Lian Wang; software, Bei Cui, Lian Wang, Quoc Khang Tran and Dinh Duy Vu; validation, Thi Tuyet Xuan Bui and Thi Tuyen Phung; investigation, Bei Cui, Lian Wang, Thi Tuyen Phung, Thi Tuyet Xuan Bui, Quoc Khang Tran, Hieu Yen Ha, Van Thang Bui and Dinh Duy Vu; resources, Thi Tuyen Phung and Dinh Duy Vu; data curation, Bei Cui, Syed Noor

Muhammad Shah and Lian Wang; writing-original draft preparation, Bei Cui, Dinh Duy Vu, Quoc Khang Tran and Syed Noor Muhammad Shah; writing-review and editing, Bei Cui, Dinh Duy Vu and Syed Noor Muhammad Shah; supervision, Lian Wang and Dinh Duy Vu; project administration, Bei Cui. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Ahmad A, Wang J, Pan Y, Sharif R, Gao S (2018) Development and use of Simple Sequence Repeats (SSRs) markers for sugarcane breeding and genetic studies. Agronomy 8 (11): 260. https://doi.org/10.3390/agronomy8110260
- Apweiler R (2004) UniProt: the Universal Protein knowledgebase. Nucleic Acids Research 32: D115-9. https://doi.org/10.1093/nar/gkh131
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The gene ontology onsortium. Nature genetics 25 (1): 25-9. https://doi.org/10.1038/75556
- Beech E, Barstow M, Rivers M (2017) The Red List of Theaceae. Botanic Gardens Conservation International Descanso House, 199 Kew Road, Richmond, Surrey, TW9 3BW, UK, 48 pp.
- Beier S, Thiel T, Münch T, Scholz U, Mascher M (2017) MISA-web: a web server for microsatellite prediction. Bioinformatics 33 (16): 2583-2585. https://doi.org/10.1093/bioinformatics/btx198
- Bosse M, van Loon S (2022) Challenges in quantifying genome erosion for conservation. Frontiers in Genetics 13: 960958. https://doi.org/10.3389/fgene.2022.960958
- Bui X, Vu D (2024) Population genetics analysis of *Diospyros mun* A.Chev. ex Lecomte (Ebenaceae) based on EST-SSR markers derived from a novel transcriptome. Biodiversity Data Journal 12: e130385. https://doi.org/10.3897/bdj.12.e130385
- Charlesworth B, Sniegowski P, Stephan W (1994) The evolutionary dynamics of repetitive DNA in eukaryotes. Nature 371 (6494): 215-220. https://doi.org/10.1038/371215a0
- Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21 (18): 3674-3676. https://doi.org/10.1093/bioinformatics/bti610
- Davey J, Hohenlohe P, Etter P, Boone J, Catchen J, Blaxter M (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics 12 (7): 499-510. https://doi.org/10.1038/nrg3012
- Deng YY, Li JQ, Wu SF, Zhu YP, Cai YW, He FC (2006) Integrated NR database in protein annotation system and its localization. Computer Engineering 32: 71-72.

- Earl D, vonHoldt B (2011) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resources 4 (2): 359-361. https://doi.org/10.1007/s12686-011-9548-7
- Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. Molecular Ecology 14 (8): 2611-2620. https://doi.org/10.1111/j.1365-294x.2005.02553.x
- Excoffier L, Laval G, Schneider S (2005) Arlequin (version 3.0): An integrated software package for population genetics data analysis. Evolutionary Bioinformatics 1: 47-50. https://doi.org/10.1177/117693430500100003
- Fan P, Zhu G, Wambulwa M, Milne R, Wu Z, Luo Y, Shahi Shavvon R, Jump A, Maity D, Gao L, Qi H, Wu H, Kong X, Khan R, Yan L, Turuspekov Y, Li D, Liu J (2025) Genetic origins and climate-induced erosion in economically important Asian walnuts. Conservation Biology 40 (1): e70125. https://doi.org/10.1111/cobi.70125
- Finn R, Mistry J, Tate J, Coggill P, Heger A, Pollington J, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy S, Bateman A (2009) The Pfam protein families database. Nucleic Acids Research 38 https://doi.org/10.1093/nar/gkp985
- Gao J, Parks CR, Du Y, (2005) Collected species of the genus *Camellia*: An illustrated outline. Zhejiang Science and Technology Press, Hangzhou [In Chinese].
- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology 29 (7): 644-652. https://doi.org/10.1038/nbt.1883
- Ha BH, Pham MP, Nguyen QK, Bui TTX, Vu G, Shah SNM, Vu DD (2025) Transcriptome characterisation and population genetics of *Cunninghamia konishii* Hayata – An endangered gymnosperm and implication for its conservation in Vietnam. Biodiversity Data Journal 13: e153663. https://doi.org/10.3897/bdj.13.e153663
- Hamrick JL, Godt MJ (1990) Plant population genetics, breeding, and genetic resources. In: Brown AHD, Clegg MT, Kahler AL, Weir BS (Eds) Allozyme diversity in plant species. Sinauer Associates Inc.
- Hedrick P (2005) A standardized genetic differentiation measure. Evolution 59 (8): 1633-1638. https://doi.org/10.1554/05-076.1
- Jakobsson M, Rosenberg N (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23 (14): 1801-1806. https://doi.org/10.1093/bioinformatics/btm233
- Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genetics 11 (1): 94. https://doi.org/10.1186/1471-2156-11-94
- Kalinowski ST, Taper ML, Marshall TC (2007) Revising how the computer program Cervus accommodates genotyping error increases success in paternity assignment. Molecular Ecology 16 (5): 1099-1106. https://doi.org/10.1111/j.1365-294x.2007.03089.x
- Kanehisa M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Research 32: D277-80. https://doi.org/10.1093/nar/gkh063
- Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, Makarova KS, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Rogozin IB, Smirnov S, Sorokin AV, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA (2004) A comprehensive evolutionary

classification of proteins encoded in complete eukaryotic genomes. Genome Biology 5 (2). https://doi.org/10.1186/gb-2004-5-2-r7

- Kõressaar T, Lepamets M, Kaplinski L, Raime K, Andreson R, Remm M (2018) Primer3_masker: integrating masking of template sequence with primer design software. Bioinformatics 34 (11): 1937-1938. https://doi.org/10.1093/bioinformatics/bty036
- Kumar R, Das SP, Choudhury BU, Kumar A, Prakash NR, Verma R, Chakraborti M, Devi AG, Bhattacharjee B, Das R, Das B, Devi HL, Das B, Rawat S, Mishra VK (2024) Advances in genomic tools for plant breeding: harnessing DNA molecular markers, genomic selection, and genome editing. Biological Research 57 (1). https://doi.org/10.1186/s40659-024-00562-6
- Li Q, Su X, Ma H, Du K, Yang M, Chen B, Fu S, Fu T, Xiang C, Zhao Q, Xu L (2021) Development of genic SSR marker resources from RNA-seq data in *Camellia japonica* and their application in the genus *Camellia*. Scientific Reports 11: 9919. https://doi.org/10.1038/s41598-021-89350-w
- Li SG, Chen SY, Fu YP, Yin AP, Sima YK, Qi RP, Zhou Y, Wang SC (2024) Genetic diversity and genetic structure of *Kadsura coccinea* germplasm resources revealed by SSR markers. Journal of Central South University of Forestry & Technology 44: 156-165. [In Chinese].
- Liu J, Möller M, Provan J, Gao L, Poudel RC, Li D (2013) Geological and ecological factors drive cryptic speciation of yews in a biodiversity hotspot. New Phytologist 199 (4): 1093-1108. https://doi.org/10.1111/nph.12336
- Li X, Wang J, Fan Z, Li J, Yin H (2019) Genetic diversity in the endangered *Camellia nitidissima* assessed using transcriptome-based SSR markers. Trees 34 (2): 543-552. https://doi.org/10.1007/s00468-019-01935-1
- Mao J, Huang D, Wang K, Peng H, Yao X, Mao Y, Jiao L, Wang H, Long Y, Tan R, Tsering O, Wang W, Tsering W, Chen L, Chen X, Leng Y (2025) Genetic diversity and population structure of tea (*Camellia sinensis*) germplasm from the Xizang Plateau. Horticulturae 12 (1): 50. https://doi.org/10.3390/horticulturae12010050
- Martínez-Campos C, Lanz-Mendoza H, Cime-Castillo J, Peralta-Zaragoza Ó, Madrid-Marina V (2025) RNA through time: From the origin of life to therapeutic frontiers in transcriptomics and epitranscriptional medicine. International Journal of Molecular Sciences 26 (11): 4964. https://doi.org/10.3390/ijms26114964
- Ming TL, Bartholomew B (2007) Theaceae. In: Wu ZY, Raven PH, Hong DY (Eds.), Flora of China, Vol. 12 (Hippocastanaceae through Theaceae). Science Press, Beijing & Missouri Botanical Garden Press, St. Louis, 366-478 pp.
- Mueller U, Wolfenbarger LL (1999) AFLP genotyping and fingerprinting. Trends in Ecology & Evolution 14 (10): 389-394. https://doi.org/10.1016/s0169-5347(99)01659-6
- Nguyen DH, Luong VD, Le TT, Tran QT, Do ND, Ly NS (2020) *Camellia puhoatensis* (Sect. Archecamellia – Theaceae), a new species from Vietnam. PhytoKeys 153: 1-11. https://doi.org/10.3897/phytokeys.153.49388
- Orel G, Nguyen MC, Rivers MC (2018) *Camellia cucphuongensis*. IUCN Red List of Threatened Species https://doi.org/10.2305/iucn.uk.2018-1.rlts.t191422a1979357.en
- Parmar R, Seth R, Sharma RK (2022) Genome-wide identification and characterization of functionally relevant microsatellite markers from transcription factor genes of tea (*Camellia sinensis* (L.) O. Kuntze). Scientific Reports 12 (1): 201. https://doi.org/10.1038/s41598-021-03848-x

- Peakall R, Smouse P (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. Bioinformatics 28 (19): 2537-2539. https://doi.org/10.1093/bioinformatics/bts460
- Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, Tsai J, Quackenbush J (2003) TIGR Gene Indices clustering tools (TGICL): a softwaresystem for fast clustering of large EST datasets. Bioinformatics 19 (5): 651-652. https://doi.org/10.1093/bioinformatics/btg034
- Pham M, Vu DD, Bei C, Bui TTX, Vu DG, Shah SNM (2024) Characterisation of the *Cinnamomum parthenoxylon* (Jack) Meisn (Lauraceae) transcriptome using Illumina paired-end sequencing and EST-SSR markers development for population genetics. Biodiversity Data Journal 12: e123405. https://doi.org/10.3897/bdj.12.e123405
- Phan H, Thi-Lan N, Le O, Nguyen C, Trinh C, Ha Y, Shah SNM, Vu D (2026) Transcriptome characterization and population genetics of *Ludisia discolor* (Ker Gawl.) A.Rich (Orchidaceae): implication for its conservation in Vietnam. Biodiversity Data Journal 14: e173579. https://doi.org/10.3897/bdj.14.e173579
- Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155 (2): 945-959. https://doi.org/10.1093/genetics/155.2.945
- Rosenberg N (2003) distruct: a program for the graphical display of population structure. Molecular Ecology Notes 4 (1): 137-138. https://doi.org/10.1046/j.1471-8286.2003.00566.x
- Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4: 406-425. https://doi.org/10.1093/oxfordjournals.molbev.a040454
- Tamura K, Stecher G, Kumar S (2021) MEGA11: Molecular Evolutionary Genetics Analysis Version 11. Molecular Biology and Evolution 38 (7): 3022-3027. https://doi.org/10.1093/molbev/msab120
- Tan L, Wang L, Wei K, Zhang C, Wu L, Qi G, Cheng H, Zhang Q, Cui Q, Liang J (2013) Floral transcriptome sequencing for SSR marker development and linkage map construction in the Tea plant (*Camellia sinensis*). PLOS ONE 8 (11): e81611. https://doi.org/10.1371/journal.pone.0081611
- Tautz D, Renz M (1984) Simple sequences are ubiquitous repetitive components of eukaryotic genomes. Nucleic Acids Research 12 (10): 4127-4138. https://doi.org/10.1093/nar/12.10.4127
- Tian Q, Huang B, Huang J, Wang B, Dong L, Yin X, Gong C, Wen Q (2022) Microsatellite analysis and polymorphic marker development based on the full-length transcriptome of *Camellia chekiangoleosa*. Scientific Reports 12 (1): 18906. https://doi.org/10.1038/s41598-022-23333-3
- Tong Y, Gao L (2020) Development and characterization of EST- SSR markers for *Camellia reticulata*. Applications in Plant Sciences 8 (5). https://doi.org/10.1002/aps3.11348
- Tran N (1998) *Camellia cucphuongensis*: a new species of yellow *Camellia* from Vietnam. International Camellia Journal 30: 71-72.
- van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004) micro-checker: software for identifying and correcting genotyping errors in microsatellite data. Molecular Ecology Notes 4 (3): 535-538. https://doi.org/10.1111/j.1471-8286.2004.00684.x

- Vieira MLC, Santini L, Diniz AL, Munhoz CdF (2016) Microsatellite markers: what they mean and why they are so useful. Genetics and Molecular Biology 39 (3): 312-328. https://doi.org/10.1590/1678-4685-gmb-2016-0027
- Vu DD, Shah SNM, Pham MP, Bui VT, Nguyen MT, Nguyen TPT (2020) De novo assembly and Transcriptome characterization of an endemic species of Vietnam, *Panax vietnamensis* Ha et Grushv., including the development of EST-SSR markers for population genetics. BMC Plant Biology 20 (1): 358. https://doi.org/10.1186/s12870-020-02571-5
- Wang F, Cheng X, Cheng S, Li W, Huang X (2023) Genetic diversity of the wild ancient tea tree (*Camellia taliensis*) populations at different altitudes in Qianjiazhai. PLOS One 18 (4): e0283189. https://doi.org/10.1371/journal.pone.0283189
- Wang R, Comptom SG, Chen XY (2011) Fragmentation can increase spatial genetic structure without decreasing pollen-mediated gene flow in a wind-pollinated tree. Molecular Ecology 20 (21): 4421-4432. https://doi.org/10.1111/j.1365-294x.2011.05293.x
- Wariss HM, Liu T, Zhang H, Wu J, Yang Z, Li W (2025) Genetic diversity and population structure of the endangered medicinal plant *Ferula sinkiangensis*. Global Ecology and Conservation 58: e03437. https://doi.org/10.1016/j.gecco.2025.e03437
- Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution 38 (6): 1358-1370. https://doi.org/10.2307/2408641
- Xia E, Jiang J, Huang H, Zhang L, Zhang H, Gao L (2014) Transcriptome analysis of the oil-rich tea plant, *Camellia oleifera*, reveals candidate genes related to lipid metabolism. PLOS One 9 (8). https://doi.org/10.1371/journal.pone.0104150
- Xie Y, Su M, Gao H, Yan G, Li S, Chen J, Bai Y, Deng J (2025) SSR marker-based genetic diversity and structure analyses of Camellia nitidissima var. phaeopubisperma from different populations. PeerJ 13 https://doi.org/10.7717/peerj.18845
- Xin T, Huang W, De Riek J, Zhang S, Ahmed S, Van Huylenbroeck J, Long C (2017) Genetic diversity, population structure, and traditional culture of (*Camellia reticulata*). Ecology and Evolution 7 (21): 8915-8926. https://doi.org/10.1002/ece3.3340
- Yang G, Ma T, Wang Z (2025) Genetic diversity and population structure of *Sphaeropteris brunoniana* based on SSR molecular markers. Global Ecology and Conservation 62: e03720. https://doi.org/10.1016/j.gecco.2025.e03720
- Ye Z, Wu Y, Ul Haq Muhammad Z, Yan W, Yu J, Zhang J, Yao G, Hu X (2020) Complementary transcriptome and proteome profiling in the mature seeds of *Camellia oleifera* from Hainan Island. PLOS One 15 (2): e0226888. https://doi.org/10.1371/journal.pone.0226888
- Zhang H, Xia E, Huang H, Jiang J, Liu B, Gao L (2015) De novo transcriptome assembly of the wild relative of tea tree (*Camellia taliensis*) and comparative analysis with tea transcriptome identified putative genes associated with tea quality and stress response. BMC Genomics 16 (1): 298. https://doi.org/10.1186/s12864-015-1494-4
- Zhao D, Parnell JN, Hodkinson T (2017) Typification of names in the genus *Camellia* (Theaceae). Phytotaxa 292 (2): 171-179. https://doi.org/10.11646/phytotaxa.292.2.4

Supplementary materials

Suppl. material 1: Species distribution

Authors: Bei Cui, Lian Wang, Quoc Khang Tran, Thi Tuyet Xuan Bui, Thi Tuyen Phung, Syed Noor Muhammad Shah, Hieu Yen Ha, Van Thang Bui, Dinh Duy Vu
Data type: images
Brief description: Species distribution of Nr-annotated unigenes in *C. cucphuongensis*.
Download file (98.66 kb)

Suppl. material 2: Frequency distribution of EST-SSR repeat motif types

Authors: Bei Cui, Lian Wang, Quoc Khang Tran, Thi Tuyet Xuan Bui, Thi Tuyen Phung, Syed Noor Muhammad Shah, Hieu Yen Ha, Van Thang Bui, Dinh Duy Vu
Data type: images
Brief description: Frequency distribution of EST-SSR repeat motif types (mono- to hexa-nucleotide repeats) identified from the transcriptome of *C. cucphuongensis*.
Download file (27.78 kb)

Suppl. material 3: Frequency distribution of SSRs

Authors: Bei Cui, Lian Wang, Quoc Khang Tran, Thi Tuyet Xuan Bui, Thi Tuyen Phung, Syed Noor Muhammad Shah, Hieu Yen Ha, Van Thang Bui, Dinh Duy Vu;
Data type: Suppl. material 1
Brief description: Frequency distribution of SSRs based on motif types in *C. cucphuongensis* transcriptome.
Download file (32.24 kb)

Suppl. material 4: Frequency of SSRs based on repeat types

Authors: Bei Cui, Lian Wang, Quoc Khang Tran, Thi Tuyet Xuan Bui, Thi Tuyen Phung, Syed Noor Muhammad Shah, Hieu Yen Ha, Van Thang Bui, Dinh Duy Vu
Data type: etc.
Brief description: Frequency of SSRs based on repeat types in *C. cucphuongensis* transcriptome.
Download file (17.39 kb)

Suppl. material 5: Pairwise genetic differentiation

Authors: Bei Cui, Lian Wang, Quoc Khang Tran, Thi Tuyet Xuan Bui, Thi Tuyen Phung, Syed Noor Muhammad Shah, Hieu Yen Ha, Van Thang Bui, Dinh Duy Vu
Data type: etc.
Brief description: Pairwise genetic differentiation (F_{ST}) and *Nm* between populations for *C. cucphuongensis* species.
Download file (13.27 kb)